Malaria Journal

**Open Access**

CrossMark

# Genomic signatures of population decline in the malaria mosquito *Anopheles gambiae*

Samantha M. O'Loughlin[1*], Stephen M. Magesa[2,3], Charles Mbogo[4,5], Franklin Mosha[6], Janet Midega[4,7,8] and Austin Burt[1]

## Abstract

**Background:** Population genomic features such as nucleotide diversity and linkage disequilibrium are expected to be strongly shaped by changes in population size, and might therefore be useful for monitoring the success of a control campaign. In the Kilifi district of Kenya, there has been a marked decline in the abundance of the malaria vector *Anopheles gambiae* subsequent to the rollout of insecticide-treated bed nets.

**Methods:** To investigate whether this decline left a detectable population genomic signature, simulations were performed to compare the effect of population crashes on nucleotide diversity, Tajima's D, and linkage disequilibrium (as measured by the population recombination parameter ρ). Linkage disequilibrium and ρ were estimated for *An. gambiae* from Kilifi, and compared them to values for *Anopheles arabiensis* and *Anopheles merus* at the same location, and for *An. gambiae* in a location 200 km from Kilifi.

**Results:** In the first simulations ρ changed more rapidly after a population crash than the other statistics, and therefore is a more sensitive indicator of recent population decline. In the empirical data, linkage disequilibrium extends 100–1000 times further, and ρ is 100–1000 times smaller, for the Kilifi population of *An. gambiae* than for any of the other populations. There were also significant runs of homozygosity in many of the individual *An. gambiae* mosquitoes from Kilifi.

**Conclusions:** These results support the hypothesis that the recent decline in *An. gambiae* was driven by the rollout of bed nets. Measuring population genomic parameters in a small sample of individuals before, during and after vector or pest control may be a valuable method of tracking the effectiveness of interventions.

**Keywords:** *Anopheles gambiae*, Kilifi, Linkage disequilibrium, Population recombination, Population control, Pest management

## Background

Many population genomic parameters depend upon population size. These include genetic diversity (θ or π), linkage disequilibrium, the population recombination parameter ρ, and runs of homozygosity [1–4]. Changes in population size can also lead to transient changes in the allele frequency spectrum and statistics based on it (e.g.

Tajima's D; [5, 6]). Thus population genomic data contain information on the past demographic history and might, therefore, show the effects of efforts to suppress the population, as has been observed in the *Plasmodium* genome after malaria control [7, 8].

In many parts of Africa there have been concerted efforts to control malaria transmission by controlling the mosquito vector using insecticide treated bed nets (ITNs) and indoor residual spraying (IRS). In some places these efforts have been successful, resulting in substantial reductions in the numbers of *Anopheles gambiae* (the most important vector species in sub-Saharan Africa)

*Correspondence: s.oloughlin@imperial.ac.uk
[1] Department of Life Sciences, Imperial College London, Silwood Park, Ascot SL5 7PY, UK
Full list of author information is available at the end of the article

O'Loughlin *et al. Malar J* (2016) 15:182

Page 2 of 10

and in malaria transmission [9, 10]. *An. gambiae* is particularly susceptible to ITNs and IRS because of its propensity to bite and rest indoors. Other vectors may be less susceptible to these control methods, including the sibling species *Anopheles arabiensis*, which is able to bite earlier and outdoors due to increased resistance to desiccation [10, 11].

One place where there has been a particularly striking reduction in *An. gambiae* abundance is in the Kilifi district of coastal Kenya: entomological surveys have revealed an overall reduction in density of *An. gambiae s.l.* and *Anopheles funestus*, accompanied by a shift in the proportions of different species, with *An. arabiensis* and *Anopheles merus* replacing *An. gambiae s.s.* and *An. funestus* as the major vectors [12]. The authors attribute this shift in species composition to the widespread distribution and use of ITNs from 2006 onwards, although they do not rule out land-use change and improvements in house construction as contributing factors.

A previous study by O'Loughlin et al. [13] reported a RADseq analysis of the *An. gambiae s.l.* species complex from three locations approx. 200 km apart in East Africa: Moshi, Muheza, and Kilifi. Although mosquito control is also being carried out in Muheza and Moshi, at the time of sampling there had been no reported decline in mosquito numbers. In this study it was found that genetic diversity in *An. gambiae* s.s. was slightly but significantly lower in Kilifi than in Muheza (~5 % lower π and ~15 % lower $\theta_W$ averaged across all chromosomes). Diversity did not differ among the three *An. arabiensis* populations. The study also found that *An. gambiae* in Kilifi was the only population with a positive value for Tajima's D, reflecting a deficit in low frequency polymorphisms, consistent with a recent decline in population size [14]. Modelling of the allele frequency spectra showed that *An. arabiensis* and *An. merus* fitted a simple model of modest population expansion, whereas the *An. gambiae* populations showed a more complex history of past population expansion followed by population decline. In the case of *An. gambiae* from Kilifi, the present population size was inferred to be smaller than the historical, pre-expansion size.

These results appear to be consistent with a recent population reduction for *An. gambiae* in Kilifi, perhaps due to control efforts [13]. To investigate this hypothesis more closely, here the analysis is expanded to consider linkage disequilibrium and the population recombination parameter ρ. RADseq data consists of a small fraction of the genome so is not suitable for some linkage-based methods of inferring population history such PSMC and MSMC [15, 16]. However the number of SNPs and their location throughout the genome make it ideal for calculating ρ. In the standard neutral model at equilibrium ρ

has an expected value (or is defined as) $4N_e r$, where $N_e$ is the effective population size and r is the recombination rate per base per generation. ρ is inversely related to the levels of linkage disequilibrium in a sample. It has previously been observed, both empirically and by simulation, that ρ is strongly affected by non-equilibrium demographics and selection [3, 17]. Although it is well established that ρ decreases after population bottlenecks followed by recovery (e.g. [3, 17, 18]), these studies did not explore the effect of very recent population declines without recovery, such as after successful vector control. Therefore, in this study, simulations are used to study the time-scale over which the different population genomic parameters are expected to change in response to reductions in population size.

## Methods

### Simulations: genomic signatures of successful control

To investigate the expected effect of a recent population crash on ρ, $\theta_W$, π and Tajima's D, sequences were simulated under different demographic scenarios using Hudson's ms [19]. A sample size of n = 26 was used (equivalent to 13 diploid individuals), chosen to match the sample size that were analysed with RADseq, and simulated sequences of 50 kb in length. Populations were simulated with an ancestral size of 2 million (the estimated long term $N_e$ for *An. gambiae* population from [13]), a mutation rate of $1.1 \times 10^{-9}$ per generation (estimated from divergence of *Drosophila* lineages [20] and assuming ten generations per year), and a ρ of 10 times present θ, which is the neutral expectation of ρ/θ calculated using the recombination rate for chromosome 3L [21] and is within the range of values seen at selective and demographic neutrality in *Drosophila* populations [22, 23]. Throughout the simulations parameters for the 3L chromosome arm were used, because 2L and 2R contain polymorphic inversions in *An. gambiae* and similarly 2R and 3R in *An. arabiensis*. Population crashes were simulated in which the population size after the crash was $10^{-2}$, $10^{-3}$, $10^{-4}$ or $10^{-5}$ of the ancestral population, and occurred 10, $10^2$, $10^3$ or $10^4$ generations in the past. Ms commands are given in Additional file 1.

### Empirical data from East Africa

Details of mosquito collections, RADseq analysis and SNP genotyping are given in [13]. Briefly, *An. gambiae*, *An. arabiensis* and *An. merus* were collected in Kilifi (Kenya), *An. gambiae* and *An. arabiensis* from Muheza (Tanzania, about 200 km south of Kilifi), and *An. arabiensis* from Moshi (Tanzania, about 200 km south-west of Kilifi). *An. gambiae* and *An. arabiensis* were collected in May–July 2010; *An. merus* in Oct 2009 and May 2010. Villages were of similar size, and house construction

O'Loughlin *et al. Malar J* (2016) 15:182

Page 3 of 10

was the same in all locations. Genomic DNA was used for RADseq analysis with the SbfI enzyme (Floragenex, Oregon). RADseq reads were aligned to the *An. gambiae* PEST reference genome using BWA v0.5.9 [24] and SNPs were called in SAMtools v0.1.18 [25]. Summary genetic diversity statistics for each species and sampling location were reported in [13]. Metrics of data used are given in Table 1.

$r^2$ was calculated for SNPs in Haploview v4.2. The expected value of $r^2$ is a function of the recombination rate and effective population size ($N_e$), but is also affected by sample size [2], so the $r^2$ was adjusted for the sample size component by subtracting $1/n$ (see Eq. 1 in [2], n = number of diploid individuals). ρ was estimated for SNPs using the composite likelihood method implemented in LDhat [26]. Under the standard neutral model ρ/θ ≈ r/μ, so the expectation of ρ/θ at neutrality can be calculated using estimated recombination rates for each *An. gambiae* chromosome arm taken from [21] and [27], and a mutation rate (μ) of $1.1 \times 10^{-9}$ per generation [20]. To investigate what combinations of severity and timing of decline are consistent with observed ρ values, further simulations were carried out, simulating alternative demographic scenarios using ms [19]. As before, 50 kb segments were simulated, with n = 26, $\mu = 1.1 \times 10^{-9}$ and ρ set at 10 × current θ, but this time the ancestral population size was adjusted such that final value of π was as close as possible to that observed in the Kilifi *An. gambiae* population (π = 0.0081). To do this the standard recursion equation for π was used (e.g. [28]).

$$\pi_t = \pi_{t-1}\left(1 - \frac{1}{2N_2}\right) + 2\mu(1 - \pi_{t-1})$$

with the initial value $\pi_0 = 4N_1\mu / (1 + 4N_1\mu)$ being the equilibrium nucleotide diversity before the crash, $\pi_t$ the diversity $t$ generations after the crash, $N_1$ and $N_2$ the effective population sizes before and after the crash, respectively, and $m$ the mutation rate. For particular values of the magnitude of population reduction ($N_2/N_1 = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$) and for particular values of time since the crash ($t = 10, 10^2, 10^3$, and $10^4$ generations ago) these equations were used to find the starting

value of $\pi_0$ that would give a $\pi_t$ close to the observed value (0.0081). These values of $N_1$, $N_2$, and $t$ were then used to generate sequences using ms [19]. 10 simulations of each scenario were performed. The ms output was converted to diploid format for estimating ρ in LDhat (to simulate 13 diploid samples). A regression analysis was then performed, of ρ as the dependent variable and crash size and generations in the past as independent variables.

Another potential sign of population crash is extended runs of homozygosity caused by mating between related individuals [4, 29]. To look for runs of homozygosity, heterozygosity at every site was plotted across the genome for each mosquito individually.

## Results

### Simulations: genomic signatures of successful control

The population genetic parameters from the simulated data sets are shown in Fig. 1. Strong population decline a long time in the past gave too few segregating sites for analyses (due to a long-term low $N_e$), so some curves are truncated. As expected, population declines lead to reductions in π, $θ_W$ and ρ, and increases in Tajima's D, with larger reductions in population size producing larger and more rapid effects on the population genomic statistics. (Simulating reductions even further in the past than 10,000 generations resulted in Tajimas's D returning towards zero (results not shown); this reflects the fact that the population has remained small but stable for a long enough time equilibrium to be restored.) Importantly, for all scenarios, ρ responded faster and/or by a larger relative amount than the other three statistics, suggesting it can be a more sensitive measure of population control.
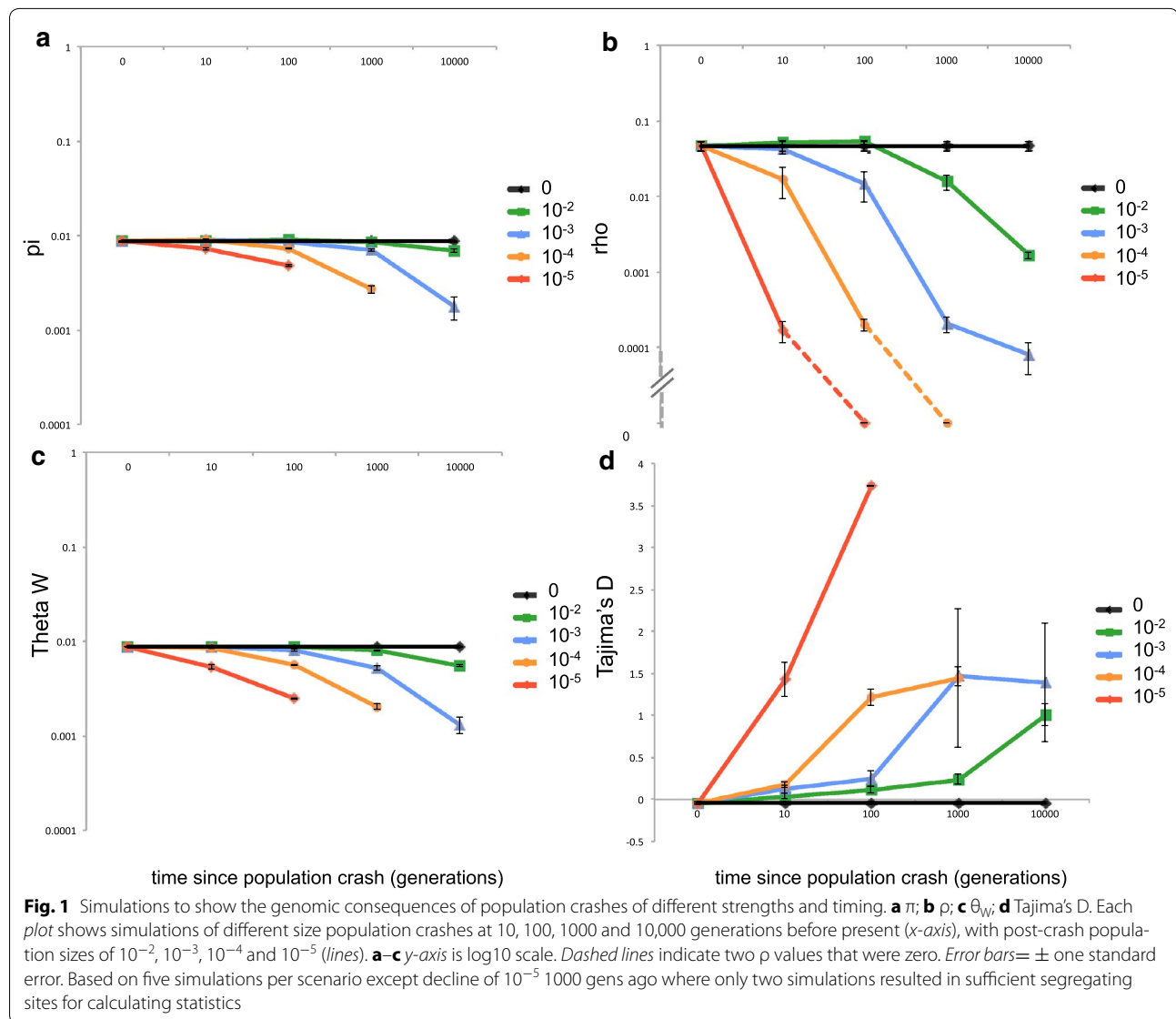
### Empirical data from East Africa

#### Linkage disequilibrium

To compare linkage disequilibrium among populations and species, the analysis was restricted to the autosomes, and because inversions greatly increase linkage disequilibrium (see Additional file 2), SNPs in the segregating inversions in the samples (2La, 2Rb and 3Ra) were removed from all calculations in all species whether the

## Table 1  Metrics of data sets containing variant and invariant sites

| Species | No. indivs. | No. of tag locations | No. bp | Mean coverage | Mean dist between RADseq tags (kb) | No. of SNPs[a] | Mean (Max) SNPs per tag |
|---------|-------------|----------------------|--------|---------------|-------------------------------------|----------------|-------------------------|
| *An. gambiae* | 24 | 2033 | 172,655 | 88 | 131 | 6387 (6437) | 3.14 (17) |
| *An. arabiensis* | 36 | 2049 | 129,315 | 106 | 130 | 5137 (5195) | 2.51 (23) |
| *An. merus* | 12 | 2280 | 253,043 | 134 | 115 | 4135 (4161) | 1.81 (13) |

[a] Number in brackets shows number of SNPs in variant only data sets. Bayesian variant calling resulted in slightly different numbers of SNPs in variant only vs genotype data sets
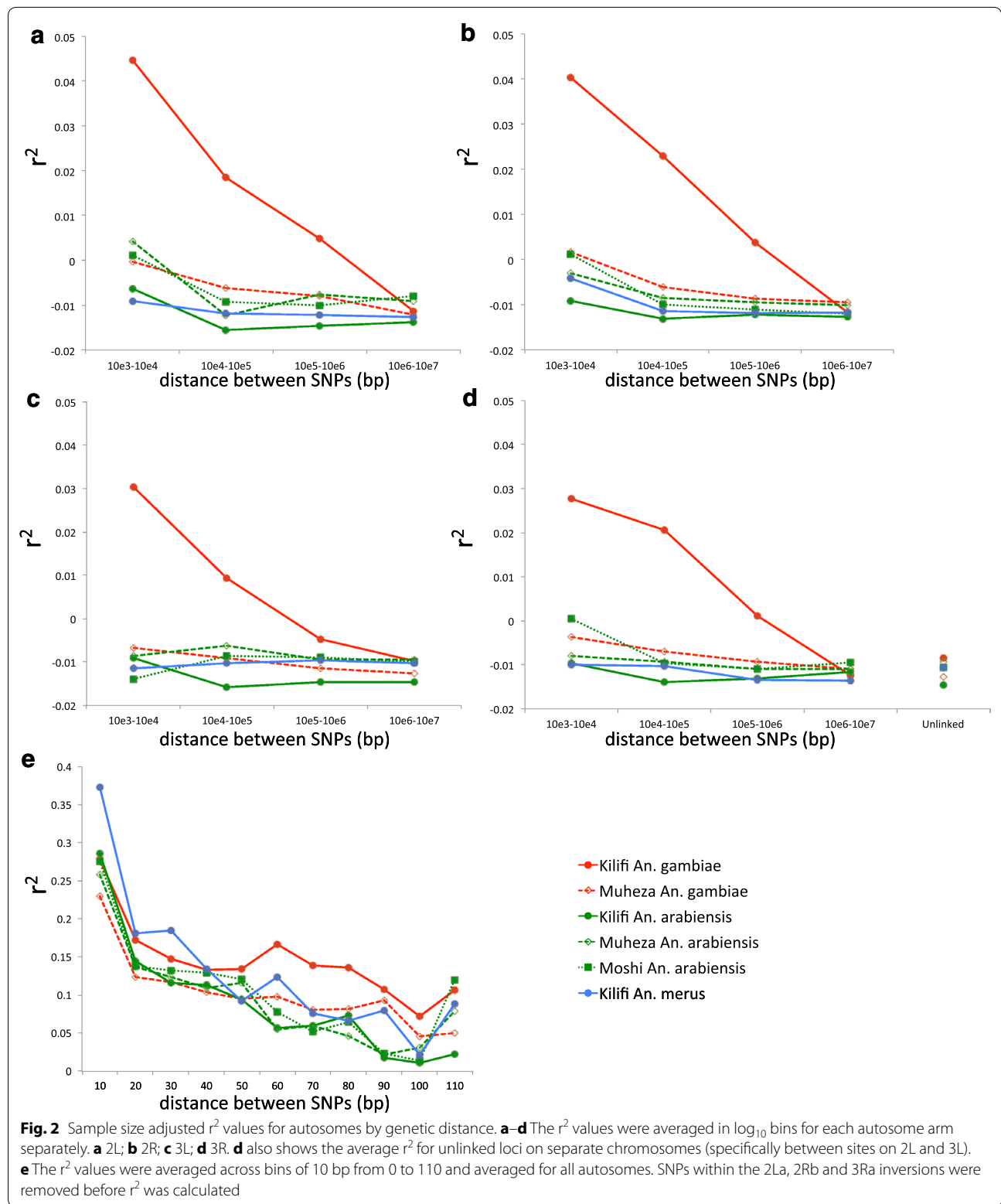
O'Loughlin *et al. Malar J* (2016) 15:182

Page 4 of 10



**Fig. 1** Simulations to show the genomic consequences of population crashes of different strengths and timing. **a** π; **b** ρ; **c** θ$_W$; **d** Tajima's D. Each *plot* shows simulations of different size population crashes at 10, 100, 1000 and 10,000 generations before present (*x-axis*), with post-crash population sizes of 10$^{-2}$, 10$^{-3}$, 10$^{-4}$ and 10$^{-5}$ (*lines*). **a**–**c** *y-axis* is log10 scale. *Dashed lines* indicate two ρ values that were zero. *Error bars*= ± one standard error. Based on five simulations per scenario except decline of 10$^{-5}$ 1000 gens ago where only two simulations resulted in sufficient segregating sites for calculating statistics

inversion was segregating or not, to ensure only collinear regions of the genome were used. For each chromosome arm, pairwise linkage disequilibrium (r$^2$) between polymorphic sites in different RADseq tag locations was calculated and averaged in bins from 10$^3$ to 10$^7$ bp (Fig. 2). For all four chromosome arms, linkage disequilibrium is substantially higher in the Kilifi population of *An. gambiae* than in any of the other populations. In these other populations linkage disequilibrium is at background levels (i.e., equal to that between polymorphisms on different chromosomes – Fig. 2d) at distances of 1–10 kb or greater, whereas for *An. gambiae* from Kilifi this only occurs at distances greater than 1 Mb. Linkage disequilibrium between polymorphisms in the same RADseq tag location (length 110 bp) were also analysed, though there are many fewer such pairs of polymorphisms and so data

was combined across the chromosomal arms (Fig. 2e). There is some tendency for linkage disequilibrium to be higher in Kilifi *An. gambiae* than in other populations, particularly for markers separated by 50–110 bp, though the difference is not as dramatic.

### Population recombination parameter ρ

In Fig. 3, estimates of ρ and ρ/θ$_W$ for each species, sampling location and chromosome arm are shown. For ease of comparison, θ$_W$ and Tajima's D from [13] are also given. Setting aside the population of *An. gambiae* from Kilifi, one can see that ρ is relatively consistent across species, populations and chromosome arms, with the exception of 2L for *An. gambiae* in Muheza and 2R for *An. arabiensis* in Moshi, for both of which ρ is much reduced (Fig. 3a). Both these samples have

O'Loughlin *et al. Malar J (2016) 15:182*

Page 5 of 10



**Fig. 2** Sample size adjusted $r^2$ values for autosomes by genetic distance. **a**–**d** The $r^2$ values were averaged in $\log_{10}$ bins for each autosome arm separately. **a** 2L; **b** 2R; **c** 3L; **d** 3R. **d** also shows the average $r^2$ for unlinked loci on separate chromosomes (specifically between sites on 2L and 3L). **e** The $r^2$ values were averaged across bins of 10 bp from 0 to 110 and averaged for all autosomes. SNPs within the 2La, 2Rb and 3Ra inversions were removed before $r^2$ was calculated

segregating inversions in the relevant arm which presumably accounts for the reduction, though it is not clear why the effect is not seen for 2R in the other populations of

*An. arabiensis.* ρ is also somewhat reduced in 2R for *An. gambiae* in Muheza; there is an inversion segregating on this arm in East African *An. gambiae*, which, even if not

O'Loughlin *et al. Malar J* (2016) 15:182

Page 6 of 10

segregating in these samples, would reduce recombination and could account for the lower value. Apart from these exceptions, the ratio $\rho/\theta_W$ is close to its neutral expectation (=r/u), though consistently somewhat lower (Fig. 3b), perhaps because r has been over-estimated, or u under-estimated, or because of population structure [30].

The results from the Kilifi population of *An. gambiae* are markedly different, with ρ low in all arms. Comparison with the Muheza population demonstrates a statistically significant difference (paired t test, t = 2.39, p = 0.04). The reduction in ρ in Kilifi vs Muheza is between 11 % (2L) and 99.9 % (3R), with an average across chromosomes of 82 %. This difference is substantially greater than the reductions seen in diversity measures π and $\theta_W$. As a result, the ratio $\rho/\theta_W$ for *An. gambiae* from Kilifi is more than 1000-fold lower then neutral expectations. The comparison among populations and species is perhaps clearest for chromosome arm 3L, which does not have segregating inversions in any of these three mosquito species (hatched bars in Fig. 3).

### Estimating the timing and severity of population crash from ρ

To investigate the combination of timing and extent of population decline consistent with the observed value of ρ for Kilifi *An. gambiae*, population crashes of varying magnitude and at varying times in the past were simulated, in each case keeping the final value of π as close as possible to that currently observed in this population (π = 0.0081). Simulations of a population decline of $10^{-5}$ more than 100 generations in the past gave highly variable results and so were excluded from further analysis. For the remaining simulations a linear regression model was fitted using Fit Model in JMP v12.0.1 (SAS Institute Inc., Cary, NC). The best fitting model was

$$\mathrm{Log10}\,(\rho) = 1.977 - 0.983 \times \mathrm{Log10}\,(N_1 N_2) - 0.717 \\ \times \mathrm{Log10}\,(\mathrm{Generations})$$

where ρ = ρ per base + 1/50,000 (to remove zero values), Generations = generations since population crash, $N_1$ = pre-crash population size, $N_2$ = post-crash population size. The model $r^2$ was 0.884. A contour plot of the model is shown in Fig. 4, which also shows the contour line corresponding to the ρ value from Kilifi gambiae 3L [with the same transformation as the simulations; Log10 (ρ per base + 1/50,000)]. The ρ value for Kilifi is compatible with a range of population crash sizes from $5 \times 10^{-4}$ to $4 \times 10^{-6}$ corresponding to times in the past

from 10,000 generations ago to ten generations ago. The first distribution of bed nets in Kilifi occurred 17 years before the mosquitoes used in this study were collected, which equates to ~170 generations. According to the model, if the crash occurred between 10 and 170 generations ago, this would correspond to a large crash of between $3 \times 10^{-5}$ (at 170 generations ago) and $4 \times 10^{-6}$ (at 10 generations ago) of the ancestral population. In terms of a reduction in mosquito numbers, a reduction of $3 \times 10^{-5}$ would mean that a starting population of one million mosquitoes per square km would be reduced to just 30 mosquitoes per square km.

### Runs of homozygosity

At least one long run of homozygosity (>4.5 Mb) was observed in 7 of the 13 Kilifi *An. gambiae* individuals, despite having RADseq data instead of full genome sequences. Examples of these are shown in Fig. 5. The runs of homozygosity in different individuals are in different genomic locations, and so cannot be attributed to a selective sweep. *Anopheles gambiae* from Muheza and all other species and populations do not show any such runs of homozygosity.
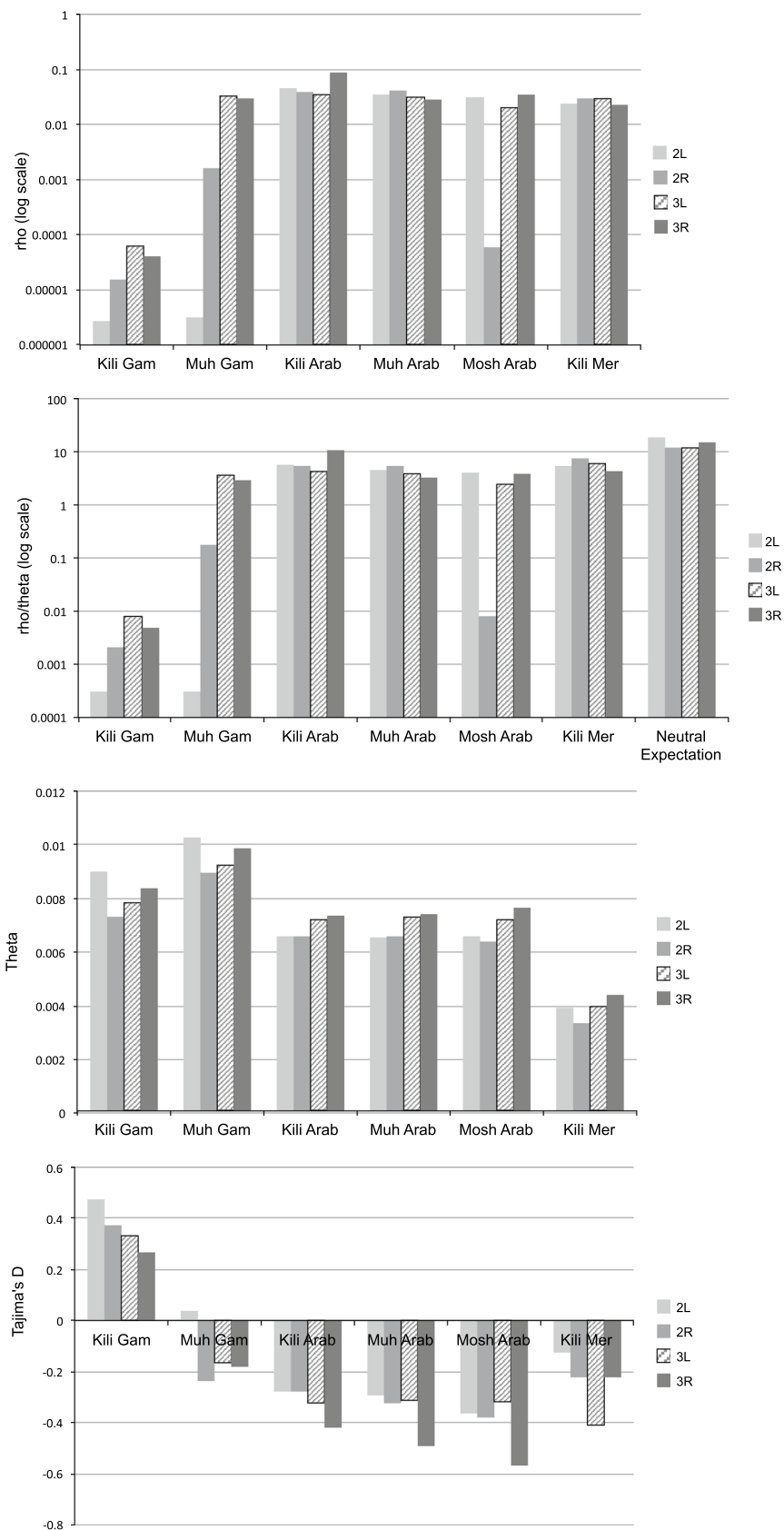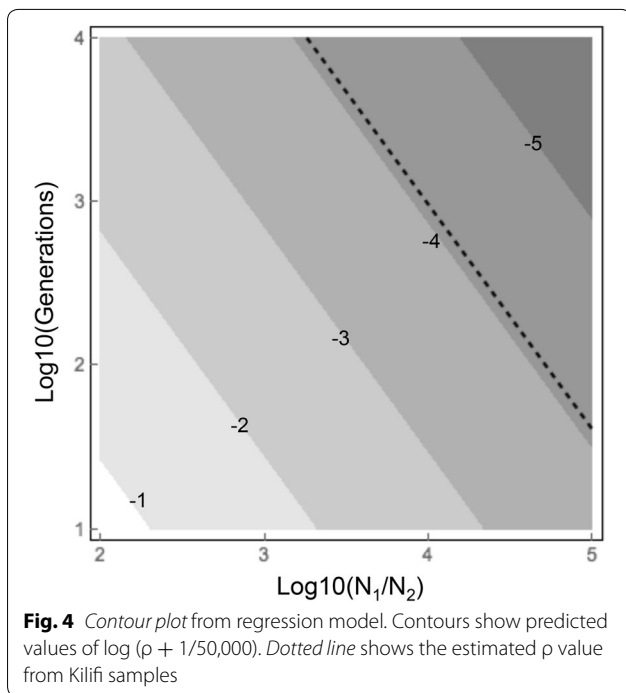
## Discussion

Vector control is an important tool in the fight against malaria and other vector borne diseases, but as yet measuring the entomological impact of these methods has been largely anecdotal or ad-hoc, with even large shifts in abundance or species proportions being difficult to quantify. This is because monitoring changes in mosquito numbers is not an easy activity; entomological surveying is labour-intensive and expensive, and prone to variation caused by seasonal fluctuations, different collection methods and degree of collection effort (e.g. [31, 32]). One alternative for detecting changes in population size may be monitoring the genome.

In the Kilifi district of Kenya, ITNs were first introduced in a randomized trial in 1993, when their effectiveness in reducing malaria incidence in children was proven [33]. Since then, ITN coverage increased gradually until a large-scale distribution program in 2006 resulted in coverage rising to 67 % [34]. Over a similar time period, entomological surveys in the Kilifi region have detected a decline in vector density (from 1990 to 2010), with an accompanying change in *An. gambiae s.l.* species composition [12]. *Anopheles gambiae* declined from 79 % of *An. gambiae s.l.* in 1997–1998 to an undetectable level

---

(See figure on next page.)

**Fig. 3** Empirical data: $\theta_W$, Tajima's D, ρ and $\rho/\theta_W$ for each species, population and autosome. ρ and $\theta_W$ are per base. Note the log scale on the *y-axis* for ρ and $\rho/\theta_W$. Inversion-less chromosome 3L is hatched for emphasis. Tajima's D and $\theta_W$ from [13]

O'Loughlin *et al. Malar J* (2016) 15:182

Page 7 of 10

O'Loughlin *et al. Malar J (2016) 15:182*

Page 8 of 10



**Fig. 4** *Contour plot* from regression model. Contours show predicted values of log (ρ + 1/50,000). *Dotted line* shows the estimated ρ value from Kilifi samples
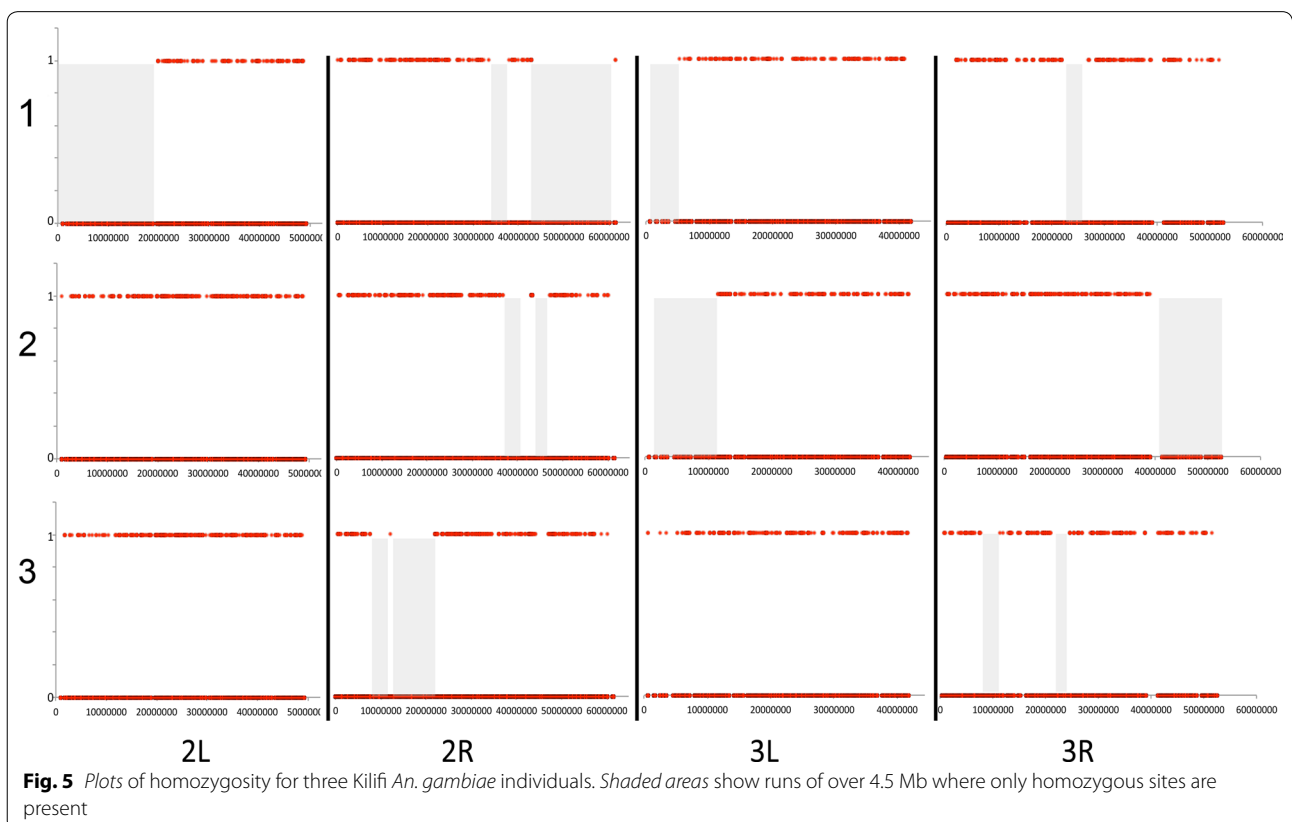
in 2007–2008. The dominant *An. gambiae s.l.* species in 2007–2008 was *An. arabiensis* (93 %), with *An. merus* contributing 5 %.

This change in mosquito abundance makes it a good model system to test whether the effect of population decline can be detected in the genome. Previously it has been shown that *An. gambiae* from Kilifi have lower π and θ and higher Tajima's D compared to those from Muheza, where there has been no reported change in mosquito abundance [13]. *Anopheles arabiensis* from the same locations did not show any differences. These results were consistent with the observed recent changes in abundance.

Here, simulations are presented showing that analysis of linkage disequilibrium and ρ can be a more sensitive test for population decline, as large signals are seen more quickly. Therefore, these statistics were compared among populations and species, and indeed linkage disequilibrium extends 100–1000 times further in Kilifi than in comparator populations, and ρ is 100–1000 times lower. Population genetic simulations indicate that the observed ρ value implies that the sampled population is no more than a $5 \times 10^{-4}$ (1/2000th) of the ancestral population. The observed values would also be consistent with an even greater, recent population decline. The unusual runs of homozygosity in Kilifi *An. gambiae* also suggest that a recent and severe population crash has occurred, which is resulting in signs of inbreeding in some individuals.



**Fig. 5** *Plots* of homozygosity for three Kilifi *An. gambiae* individuals. *Shaded areas* show runs of over 4.5 Mb where only homozygous sites are present

O'Loughlin *et al. Malar J* (2016) 15:182

Page 9 of 10

One weakness of the simulations is that they do not take into account spatial structure, which is known to effect the extent of linkage disequilibrium, especially when all the samples are from a single sub-population in a large "stepping-stone" array [3, 30]. In general, there is little differentiation among these East African populations of the same species [13], but since the extent of differentiation is determined by the product of the migration rate and the population size, there may be some interaction between a population crash and population structure affecting linkage disequilibrium and ρ. The inferences would also be improved by analysing a time series of samples from before and after the population crash, but unfortunately pre-control samples were not available for analysis.

The fact that a similar pattern of increased linkage disequilibrium and reduced ρ was not observed in *An. arabiensis* and *An. merus* collected from Kilifi at the same time as the *An. gambiae* samples suggests that these species have not been affected by the same population decline, and is consistent with the entomological observations of a shift in species composition [12]. This difference among species supports the hypothesis that the population decline has largely been due to the use of ITNs, as these are expected to have a larger impact on the highly anthropophilic and indoor-biting *An. gambiae*, compared with partially zoophilic and outdoor biting species such as *An. arabiensis* and *An. merus* [35].

The results of the simulations and observed data from Kilifi show promise for the prospective monitoring of vector control efforts. Mosquitoes have ~10 generations per year, so measurements of ρ, θ, π and Tajima's D taken pre-intervention and at one-yearly intervals should detect whether control is succeeding. Whole-genome sequencing is not necessary for measuring ρ. The simulated data suggests it is possible to get reliable estimates of ρ from as few as 300 segregating sites, so a medium-throughput SNP genotyping platform such as RADseq or Golden-Gate assay would be sufficient for monitoring. For *Anopheles* species it is important to use SNPs that are not on chromosome arms containing segregating inversions. The sample here of just 11–13 mosquitoes per population was sufficient to distinguish clear differences in ρ between populations and chromosome arms.

## Conclusions

Observations of genomic diversity and linkage disequilibrium in a small sample of just 13 mosquitoes provide compelling evidence that *An. gambiae* in Kilifi has undergone a recent population crash. In practical terms, this means that regular monitoring of a small number of genomes could allow rapid detection of whether a control intervention is succeeding. In nature, there may be complicating factors such as seasonal variation and immigration from non-treated areas, but the results presented here from Kilifi suggest that a severe population crash will be detectable despite these factors. Given the practical difficulties of measuring mosquito abundance by direct surveying, genotyping a small number of mosquitoes could be an attractive alternative for assessing the entomological impact of vector control.

Data available in Dryad Digital Repository: doi:10.5061/dryad.hm6tt.

## Additional files

**Additional file 1.** Ms commands used for simulations.

**Additional file 2.** Sample size adjusted LD decay curves for *An. gambiae* and *An. arabiensis* (collinear regions only, without 2La, 2Rb and 3Ra SNPs) and for *An. gambiae* 2La, *An. arabiensis* 2Rb and *An. arabiensis* 3Ra separately.

**Author details**
[1] Department of Life Sciences, Imperial College London, Silwood Park, Ascot SL5 7PY, UK. [2] NIMR Amani Research Centre, P.O. Box 81, Muheza, Tanzania. [3] Global Health Division, RTI International, Dar es Salaam, Tanzania. [4] Centre for Geographic Medicine Research Coast, Kenya Medical Research Institute, P.O. Box 428, Kilifi, Kenya. [5] Malaria Public Health Department, Centre for Geographic Medicine, KEMRI-Wellcome Trust Research Programme, Kenyatta National Hospital Grounds, P.O. Box 43640-00100, Nairobi, Kenya. [6] Kilimanjaro Christian Medical University College, Moshi, Tanzania. [7] Department of Life Sciences, Imperial College London, South Kensington Campus, London SW7 2AZ, UK. [8] Wellcome Trust Centre for Human Genetics, Oxford OX3 7BN, UK.

**References**
1. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975;7:256–76.
2. Hill WG. Estimation of effective population-size from data on linkage disequilibrium. Genet Res. 1981;38:209–16.
3. Andolfatto P, Przeworski M. A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics. 2000;156:257–68.

O'Loughlin *et al. Malar J* (2016) 15:182

Page 10 of 10

4. Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF. Genomic runs of homozygosity record population history and consanguinity. PLoS One. 2010;5:e13996. doi:10.1371/journal.pone.0013996.

5. Tajima F. Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123:585–95.

6. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5:e1000695. doi:10.1371/journal.pgen.1000695.

7. Nkhoma SC, Nair S, Al-Saai S, Ashley E, Mcgready R, Phyo AP, et al. Population genetic correlates of declining transmission in a human pathogen. Mol Ecol. 2013;22:273–85. doi:10.1111/mec.12099.

8. Daniels RF, Schaffner SF, Wenger EA, Proctor JL, Chang HH, Wong W, et al. Modeling malaria genomics reveals transmission decline and rebound in Senegal. Proc Natl Acad Sci USA. 2015;112:7067–72. doi:10.1073/pnas.1505691112.

9. Bayoh MN, Mathias DK, Odiere MR, Mutuku FM, Kamau L, Gimnig JE, et al. *Anopheles gambiae*: historical population decline associated with regional distribution of insecticide-treated bed nets in western Nyanza Province, Kenya. Malar J. 2010;9:62. doi:10.1186/1475-2875-9-62.

10. Russell TL, Govella NJ, Azizi S, Drakeley CJ, Kachur SP, Killeen GF. Increased proportions of outdoor feeding among residual malaria vector populations following increased use of insecticide-treated nets in rural Tanzania. Malar J. 2011;10:80. doi:10.1186/1475-2875-10-80.

11. Lindblade KA, Gimnig JE, Kamau L, Hawley WA, Odhiambo F, Olang G, et al. Impact of sustained use of insecticide-treated bednets on malaria vector species distribution and culicine mosquitoes. J Med Entomol. 2006;43:428–32. doi:10.1603/0022-2585.

12. Mwangangi JM, Mbogo CM, Orindi BO, Muturi EJ, Midega JT, Nzovu J, et al. Shifts in malaria vector species composition and transmission dynamics along the Kenyan coast over the past 20 years. Malar J. 2013;12:13. doi:10.1186/1475-2875-12-13.

13. O'Loughlin SM, Magesa S, Mbogo C, Mosha F, Midega J, Lomas S, et al. Genomic analyses of three malaria vectors reveals extensive shared polymorphism but contrasting population histories. Mol Biol Evol. 2014;31:889–902. doi:10.1093/molbev/msu040.

14. Tajima F. The effect of change in population-size on DNA polymorphism. Genetics. 1989;123:597–601.

15. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475:493–6. doi:10.1038/nature10231.

16. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. 2014;46:919–25. doi:10.1038/ng.3015.

17. Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. Genome Res. 2005;15:790–9. doi:10.1101/gr.3541005.

18. Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. Am J Hum Genet. 2001;69:1–14. doi:10.1086/321275.

19. Hudson RR. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics. 2002;18:337–8. doi:10.1093/Bioinformatics/18.2.337.

20. Tamura K, Subramanian S, Kumar S. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. Mol Biol Evol. 2004;21:36–44. doi:10.1093/molbev/msg236.

21. Pombi M, Stump AD, Della Torre A, Besansky NJ. Variation in recombination rate across the X chromosome of *Anopheles gambiae*. Am J Trop Med Hyg. 2006;75:901–3.

22. Bachtrog D, Andolfatto P. Selection, recombination and demographic history in *Drosophila miranda*. Genetics. 2006;174:2045–59. doi:10.1534/genetics.106.062760.

23. Thornton K, Andolfatto P. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics. 2006;172:1607–19. doi:10.1534/genetics.105.048223.

24. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60. doi:10.1093/bioinformatics/btp324.

25. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9. doi:10.1093/bioinformatics/btp352.

26. McVean G, Awadalla P, Fearnhead P. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics. 2002;160:1231–41.

27. Stump AD, Pombi M, Goeddel L, Ribeiro JMC, Wilder JA, Torre AD, et al. Genetic exchange in 2La inversion heterokaryotypes of *Anopheles gambiae*. Insect Mol Biol. 2007;16:703–9.

28. Gillespie JH. Population genetics: a concise guide. Baltimore: John Hopkins University Press; 2004.

29. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L et al. Erratum: runs of homozygosity in European populations (Am J Hum Genet. 2008;83:359-72, 2008). Am J Hum Genet. 2008;83:658. doi:10.1016/j.ajhg.2008.10.009.

30. De A, Durrett R. Stepping-stone spatial structure causes slow decay of linkage disequilibrium and shifts the site frequency spectrum. Genetics. 2007;176:969–81. doi:10.1534/genetics.107.071464.

31. Govella NJ, Chaki PP, Geissbuhler Y, Kannady K, Okumu F, Charlwood JD, et al. A new tent trap for sampling exophagic and endophagic members of the *Anopheles gambiae* complex. Malar J. 2009;8:157. doi:10.1186/1475-2875-8-157.

32. Chaki PP, Mlacha Y, Msellemu D, Muhili A, Malishee AD, Mtema ZJ, et al. An affordable, quality-assured community-based system for high-resolution entomological surveillance of vector mosquitoes that reflects human malaria infection risk patterns. Malar J. 2012;11:172. doi:10.1186/1475-2875-11-172.

33. Nevill CG, Some ES, Mungala VO, Mutemi W, New L, Marsh K, et al. Insecticide-treated bednets reduce mortality and severe morbidity from malaria among children on the Kenyan coast. Trop Med Int Health. 1996;1:139–46.

34. Okiro EA, Hay SI, Gikandi PW, Sharif SK, Noor AM, Peshu N, et al. The decline in paediatric malaria admissions on the coast of Kenya. Malar J. 2007;6:151. doi:10.1186/1475-2875-6-151.

35. Kipyab PC, Khaemba BM, Mwangangi JM, Mbogo CM. The bionomics of *Anopheles merus* (Diptera: culicidae) along the Kenyan coast. Parasit Vectors. 2013;6:37. doi:10.1186/1756-3305-6-37.